A Simple and Robust Approach to Detecting Subject-Verb Agreement Errors

Flachs, Simon; Lacroix, Ophélie; Rei, Marek; Yannakoudakis, Helen; Søgaard, Anders

Published in:
Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)

DOI:
10.18653/v1/N19-1251

Publication date:
2019

Document version
Publisher's PDF, also known as Version of record

A Simple and Robust Approach to Detecting Subject–Verb Agreement Errors

Simon Flachs[1,2], **Ophélie Lacroix**[1],
Marek Rei[3], **Helen Yannakoudakis**[3], **Anders Søgaard**[2]
[1] Siteimprove, Denmark
[2] CoAStaL DIKU, Department of Computer Science, University of Copenhagen, Denmark
[3] The ALTA Institute, Dept. of CS & Technology, University of Cambridge, United Kingdom
{sfl, ola}@siteimprove.com,
{marek.rei, helen.yannakoudakis}@cl.cam.ac.uk,
soegaard@di.ku.dk

Abstract

While rule-based detection of subject–verb agreement (SVA) errors is sensitive to syntactic parsing errors and irregularities and exceptions to the main rules, neural sequential labelers have a tendency to overfit their training data. We observe that rule-based *error generation* is less sensitive to syntactic parsing errors and irregularities than error detection and explore a simple, yet efficient approach to getting the best of both worlds: We train neural sequential labelers on the combination of large volumes of silver standard data, obtained through rule-based error generation, and gold standard data. We show that our simple protocol leads to more robust detection of SVA errors on both in-domain and out-of-domain data, as well as in the context of other errors and long-distance dependencies; and across four standard benchmarks, the induced model on average achieves a new state of the art.

1 Introduction

Grammatical Error Detection. Grammatical Error Detection (GED, Leacock et al., 2010) is the task of detecting grammatical errors in text. It is used in various real-world applications, such as writing assistance tools, self-assessment frameworks and language tutoring systems, facilitating incremental and/or exploratory editing of one's writing. Accurate error detection systems also have potential applications for language generation and machine translation systems, guiding automatically generated output towards grammatically correct sequences.

The problem of detecting subject–verb agreement (SVA) errors is an important subtask of GED. In this work, we focus on detecting subject–verb agreement errors in the English as a Second Language (ESL) domain. Most SVA errors occur at the third-person present tense when determining whether the subject describes a singular or a plural concept. The following examples demonstrate subject–verb agreement errors (bold):

(1) a. *They all **knows** where the conference is.
b. *The Hotel **are** very close to Town Hall.

The task can be formulated as a sequence labeling problem, with the goal of labeling subject–verb pairs as being in agreement or not.

Approaches. Sequence labeling problems in NLP, including GED and the subtask of identifying SVA errors, have, in recent years, been handled with Recurrent Neural Networks (RNNs) trained on large amounts of data (Rei and Yannakoudakis, 2016, 2017). However, most publicly available datasets for GED are relatively small, making it difficult to learn a general grammar representation and potentially leading to over-fitting. Previous work has also shown that neural language models with a similar architecture have difficulty learning subject–verb agreement patterns in the presence of agreement attractors (Linzen et al., 2016).

Rule-based approaches (Andersen et al., 2013) are still considered a strong alternative to end-to-end neural networks, with many industry solutions still relying on rules defined over syntactic trees. The rule-based approach has the advantage of not requiring manual annotation, while also allowing easy access to adding and removing individual rules. On the other hand, language is continuously evolving, and there are exceptions to most grammar rules we know. Additionally, rule-based matching typically relies on syntactic pre-processing, which is error-prone, leading to compounding errors that hurt the downstream GED performance.

Our contributions. In this work, we compare the performance of rule-based approaches and

Proceedings of NAACL-HLT 2019, pages 2418–2427
Minneapolis, Minnesota, June 2 - June 7, 2019.

end-to-end neural models for the detection of SVA errors. We show that rule-based systems are vulnerable to errors in the underlying syntactic parsers, while also failing to capture irregularities and exceptions. In contrast, end-to-end neural architectures are limited by the available labeled examples and sensitive to the variance in these datasets. We then make the following observation: while rule-based error *detection* is severely affected by errors and irregularities in syntactic parsing, rule-based error *generation* is more robust. SVA errors can be generated without identifying subject dependency relations in advance, and changing the number of a verb almost always leads to an error. This generated data can be used as a silver standard for optimizing neural sequence labeling models. We demonstrate that a system trained on a combination of available labeled data and large volumes of silver standard data outperforms both neural and rule-based baselines by a margin on three out of four standard benchmarks, and on average achieves a new state-of-the-art on detecting SVA errors.

2 Related work

Neural approaches. Recent neural approaches to GED include Rei and Yannakoudakis (2016) who argue that bidirectional (bi-) LSTMs, in particular, are superior to other RNNs when evaluated on standard ESL benchmarks for GED and give state-of-the-art results. Rei and Yannakoudakis (2017) show even better performance using a multi-task learning architecture for training bi-LSTMs that additionally predicts linguistic properties of words, such as their part of speech (PoS).

Recent studies (Linzen et al., 2016; Gulordava et al., 2018; Kuncoro et al., 2018) have specifically analyzed the performance of LSTMs in learning syntax-sensitive dependencies such as SVA.

Rule-based approaches. Cai et al. (2009) use a combination of dependency parsing and sentence simplification, as well as special handling of *wh*-elements, to detect SVA errors. Once the subject–verb relation is identified, after parsing the simplified input sentence, a PoS tagger is used to check agreement. This is similar in spirit to the rule-based baseline system used in our experiments below. Wang et al. (2015) use a similar approach, distinguishing between four different sentence types and using slightly different rules for each type. Their rules are, again, defined over the outputs of a dependency parser and a PoS tagger. Sun et al. (2007) use labeled data to derive rules based on dependency tree patterns.

Automatic error generation. Because of the scarcity of annotated datasets in GED, research has been carried out on creating artificial errors, where errors are injected into otherwise correct text using deterministic rules or probabilistic approaches using linguistic information (Felice and Yuan, 2014; Kasewa et al., 2018). Studies focusing on detecting specific error types such as determiners and prepositions (Rozovskaya and Roth, 2011) or noun number (Brockett et al., 2006) are mainly developed within the framework of automatic error generation. Recent work, expanding the detection (Rei et al., 2017) and the correction (Xie et al., 2018) tasks to all types of errors, improves the performance of neural models by training on additional artificial error data generated via machine translation methods.

Miscellaneous. Recent work has also led to good performance in correcting grammatical errors (Yannakoudakis et al., 2017; Bryant and Briscoe, 2018; Chollampatt and Ng, 2018). However, in this paper, we are interested in the task of grammatical error *detection* and we therefore compare our work to current state-of-the-art approaches to detecting errors and do not report the performance of correction systems.

3 Subject–verb agreement detection

Following recent work on GED (Rei and Yannakoudakis, 2016), we define SVA error detection as a sequence labeling task, where each token is simply labeled as correct or incorrect. For a given SVA error, only the verb is labeled as incorrect. Error types other than SVA are ignored, i.e., we do not correct the errors in the text and we do not attempt to predict them as incorrect.

In this paper, we only study SVA in English. We note that even for English, there is some controversy about what constitutes an SVA error. Manaster-Ramer (1987), cites this example, which has been used by some as an argument for English exhibiting cross-serial dependencies:

(2) The man and the women dance and sing, respectively.

We also note that subject–verb agreement can

be more or less pervasive across languages, depending on how rich the morphology is, whether the given language exhibits *pro-drop*, and how far apart subjects and verbs are likely to occur.

4 Systems

4.1 Rule-based system

Typically, building a GED rule-based system is time-consuming and requires specific knowledge to deal with the multiple exceptions and irregularities of languages. Difficult cases (such as long distance subject–verb relations) are often ignored in order to ensure high precision, at the expense of the recall of the system. However, our rule-based system is not limited to the detection of simple cases of SVA errors. It relies on PoS tags and dependency relations to identify all types of SVA errors. Specifically, our rule-based system operates as follows: (i) it identifies the candidate verbs based on PoS tags;[1] (ii) for a given verb, it uses the dependency relations to find its subject;[2] (iii) the PoS tag of the verb and its subject are used to check whether they agree in number and person. We use predicted Penn Treebank PoS tags and dependency relations provided by the Stanford Log-linear PoS Tagger (Toutanova et al., 2003) and the Stanford Neural Network Dependency Parser (Chen and Manning, 2014) respectively.

4.2 Neural system

We use the state-of-the-art neural sequence labeling architecture for error detection (Rei and Yannakoudakis, 2016). The model receives a sequence of tokens $(w_1, ..., w_T)$ as input and outputs a sequence of labels $(l_1, ..., l_T)$, i.e., one for each token, indicating whether a token is grammatically correct (in agreement) or not, in the given context. All tokens are first mapped to distributed word representations, pre-trained using word2vec (Mikolov et al., 2013) on the Google News corpus. Following Lample et al. (2016), character-based representations are also built for every word using a bi-LSTM (Hochreiter and Schmidhuber, 1997) and then concatenated onto the word embedding.

The combined embeddings are then given as input to a word-level bi-LSTM, creating representations that are conditioned on the context from both sides of the target word. These representations are then passed through an additional feed-forward layer, in order to combine the extracted features and map them to a more suitable space. A softmax output layer returns the probability distribution over the two possible labels (*correct* or *incorrect*) for each word. We also include the language modeling objective proposed by Rei (2017), which encourages the model to learn better representations via multi-tasking and predicting surrounding words in the sentence. Dropout (Srivastava et al., 2014) with probability 0.5 is applied to word representations and to the output from the word-level bi-LSTM. The model is optimised using categorical cross-entropy with AdaDelta (Zeiler, 2012).

5 Data

5.1 Data preprocessing

As the public datasets either have their own taxonomy or they are not annotated with error types at all, we apply the error type extraction tool of Bryant, Felice, and Briscoe (2017) to automatically get error types mapped to the same taxonomy for all datasets. The tool automatically annotates parallel original and corrected sentences with error type information. When evaluated by human raters, the predicted error types were rated as "good" or "acceptable" in at least 95% of the cases. We use their publicly available tool[3] to automatically get error types for all public datasets mapped to the same taxonomy of 25 error types in total. We then set SVA errors as our target class.

5.2 Test data

We compare the rule-based and neural approaches for the task of SVA error detection on four benchmarks in the ESL domain.

- **FCE**. The Cambridge Learner Corpus of First Certificate in English (FCE) exam scripts consists of texts produced by ESL learners taking the FCE exam, which assesses English at the upper-intermediate proficiency level (Yannakoudakis et al., 2011). We use the publicly available test set.

- **AESW**. The dataset from the Automated Evaluation of Scientific Writing Shared Task

[1] Present tense verbs + "was" and "were".

[2] The subject can be direct – attached with a `nsubj` relation – or indirect, such as when the syntactic subject is a relative pronoun, e.g., *who*, or an expletive, e.g., *there*.

[3] https://github.com/chrisjbryant/errant

2016 (AESW) is a collection of text extracts from published journal articles (mostly in physics and mathematics) along with their (sentence-aligned) corrected counterparts (Daudaravicius et al., 2016). We test on the combined trained, development and test set.[4]

- **JFLEG**. The JHU Fluency-Extended GUG corpus (JFLEG) represents a cross-section of ungrammatical data, consisting of sentences written by ESL learners with different proficiency levels and L1s (Napoles et al., 2017). We evaluate our models on the public test set.

- **CoNLL14**. The test dataset from the CoNLL 2014 shared task consists of (mostly argumentative) essays written by advanced undergraduate students from the National University of Singapore, and are annotated for grammatical errors by two native speakers of English (Ng et al., 2014).

5.3 Training data

ESL writings. We use the following ESL datasets as training data:

- **Lang8** is a parallel corpus of sentences with errors and their corrected versions created by scraping the Lang-8 website[5], which is an open platform where language learners can write texts and native speakers of that language can provide feedback via error correction (Mizumoto et al., 2011). It contains $1,047,393$ sentences.

- **NUCLE** comprises around $1,400$ essays written by students from the National University of Singapore. It is annotated for error tags and corrections by professional English instructors (Dahlmeier et al., 2013). It contains $57,151$ sentences.

- **FCE train set**. We use the publicly available FCE training set, containing $25,748$ sentences. A subset of $5,000$ sentences was separated and used for development experiments.

Artificial errors. We generate artificial subject–verb agreement errors from large amounts of data. Specifically, we use the British National Corpus (BNC, BNC-Consortium et al., 2007), a collection of British English sentences that includes samples from different media such as newspapers, journals, letters or essays. Subject–verb agreement in English merely consists of inflecting 3rd person singular verbs in the present tense (and *be* in the past), which makes any text in English fairly easy to corrupt with SVA errors. We assume that the BNC data is written in correct British English. Using predicted PoS tags provided by the Stanford Log-linear PoS Tagger, we identify verbs in present tense, as well as *was* and *were* for the past tense, and flip them to their respective opposite version using the list of inflected English words (annotated with morphological features) from the Unimorph project (Kirov et al., 2016). The final artificial training set includes the sentences with injected errors ($265,742$ sentences), their original counterpart, and sentences where SVA errors could not be injected due to not containing candidate verbs that could be flipped ($241,295$ sentences).

6 Experiments

The models. We compare our neural model trained on both artificially generated errors and ESL data ($\text{LSTM}_{ESL+art}$) to three baselines: a neural model trained only on ESL data (LSTM_{ESL}) (i.e., reflecting the performance of current state-of-the-art approaches for GED), a language model based method (BERT-LM) and our rule-based system.

In order to measure the real performance of a language model (LM) on the detection of SVA errors, we choose to use the BERT system (Devlin et al., 2018) to assign probabilities to different versions of the test sentences. Specifically, we use the pre-trained uncased BERT-Base model. We duplicate the sentences each time a corruptible verb occurs (flipping its number). The LM assigns a probability to both possible versions of the verbs. We select the version which has the highest probability, if this probability is at least 0.1[6] higher than the probability of the verb in the original sentence.

[4] Sentences containing special placeholders for mathematical equations, dates, etc. are filtered out.

[5] http://lang-8.com/

[6] We tune the threshold on the test dataset from the CoNLL 2013 shared task on Grammatical Error Correction of ESL learner essays.

	FCE			AESW			CoNLL14			JFLEG			
System	P	R	$F_{0.5}$	P	R	$F_{0.5}$	P	R	$F_{0.5}$	P	R	$F_{0.5}$	$F_{0.5}$ avg.
Rules	43.75	40.23	43.00	14.82	**49.75**	17.24	27.93	31.96	28.65	37.50	**48.21**	39.24	32.03
BERT-LM	66.67	**52.87**	63.36	18.36	39.61	20.57	50.00	35.24	46.13	60.00	32.14	51.14	45.30
LSTM$_{ESL}$	71.88	26.44	53.49	**27.75**	10.33	20.75	**54.84**	17.53	38.46	**73.91**	30.91	**57.82**	42.63
LSTM$_{ESL+art}$	**72.41**	48.84	**66.04**	19.05	40.66	**21.31**	49.32	**37.11**	**46.27**	64.71	39.29	57.29	**47.73**

Table 1: Performance of our systems (rule-based and LSTMs) and baselines. BERT-LM is the language model baseline.

Hyper-parameters. We tune the model hyper-parameters on the FCE development set, according to the $F_{0.5}$ score. Training is stopped when $F_{0.5}$ on the FCE development set does not improve over 7 epochs. Word representations have size 300, while character representations have size 100. The word-level LSTM hidden layers have size 300 for each direction, and the character-level LSTM hidden layers have size 100 for each direction.

Evaluation. Existing approaches are typically optimised for high precision at the cost of recall, as a system's utility depends strongly on the ratio of true to false positives, which has been found to be more important in terms of learning effect. A high number of false positives would mean that the system often flags correct language as incorrect, and may therefore end up doing more harm than good (Nagata and Nakatani, 2010). Because of this, $F_{0.5}$ is preferred to F_1 in the GED domain as it puts more weight on precision than recall. For each experiment, we report the token-level precision (P), the recall (R), and the $F_{0.5}$ scores.

7 Results

The main results are summarized in Table 1. Looking at the performance of the LSTM$_{ESL+art}$ system, we see that on 3 out of 4 benchmarks, our neural model trained on artificially generated errors outperforms the LSTM$_{ESL}$ system with respect to $F_{0.5}$. On average, over the four benchmarks, its $F_{0.5}$ score is 2.43 points higher than the best performing baseline. Both neural models obtain higher $F_{0.5}$ scores than the rule-based baseline, on average and across the board, i.e., $+10.6$ for LSTM$_{ESL}$ and $+15.7$ for LSTM$_{ESL+Art}$. The BERT-LM outperforms the LSTM$_{ESL}$ (mostly due to its higher recall, i.e., $+18.66$) but still does not reach the $F_{0.5}$ score of the LSTM$_{ESL+Art}$ system which gets higher precision and recall overall ($+2.62$ and $+1.51$ respectively).

Furthermore, we observe a trend that the two LSTM systems trade off precision and recall, with the LSTM$_{ESL}$ system yielding the highest precision across most datasets, but also yielding significantly lower recall than LSTM$_{ESL+Art}$. It is also evident that the performance varies over domains: all models struggle with AESW. This is likely due to the complexity of the scientific writing genre where, for example, sentences contain parentheses interposed between a verb and its subject. We also note errors are far less frequent in this genre, leading to moderate recall and very low precision. For the rest of the datasets, system performance is generally better.

8 Analysis

We analyze the effect of adding artificial errors to the training data. In particular, we focus on the robustness of our models by looking at how sensitive they are to grammatical errors in the surrounding context; and by looking at how good the models are at predicting agreement relative to the distance between the subject and verb. This set of experiments is similar in spirit to Linzen et al. (2016). We also analyze our rule-based baseline: so far, we know our rule-based baseline was sensitive to parser errors and irregularities. We inspect the quality of the underlying parser by evaluating it on data that resembles the data used in our experiments, to see whether errors seem to result more from parser errors or irregularities. Finally, we also look at the sensitivity of our systems to other linguistic phenomena such as relative clauses or conjunctions.

8.1 Sensitivity to other errors in the surrounding context

In ESL writings, multiple errors can occur in the same sentence. This means more variable contexts, which can lead to degradation in the performance of both syntactic parsers / rule-based systems and GED models.

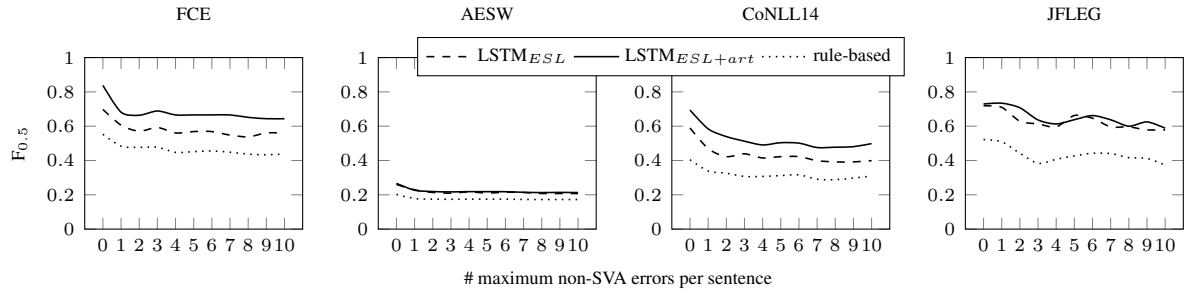

Figure 1: Performance ($F_{0.5}$ scores) of the systems with respect to the noise in test data (i.e., the number of additional non-SVA errors in sentences).

Testing on noisy contexts We first evaluate how our systems are impacted by additional non-SVA errors in the surrounding context of SVA errors in our test data. For each of the test datasets, we create multiple versions, allowing for n non-SVA errors per sentence (we correct the extra non-SVA errors). This way we can create datasets with different levels of complexity with respect to the grammatical errors within them.

In Figure 1, the $F_{0.5}$ scores of the models are shown for different numbers of grammatical errors per sentence. It is evident that all of the models are negatively affected by the presence of other errors in the same sentence. Using more data for training – i.e., our artificial training data which does not include context errors – generally boosts performance on data with and without grammatical errors in the context. In other words, training with additional artificially generated errors seems, overall, to be making our model more robust. We also note that our rule-based baseline is affected by errors to roughly the same extent as our baseline neural model is. One might have thought the rule-based baseline would suffer more, because of it being sensitive to errors in the underlying syntactic parser. We return to this issue below.

Training on non-noisy contexts In order to assess the benefit of training on non-erroneous contexts, we create a new dataset from our ESL training data (see §5.3). Based on the annotations in the data, we apply the corrections of error types other than SVA, thereby only leaving SVA errors in the data. We experiment with how adding this 'clean' dataset to the training set of our existing systems affects performance. The resulting $F_{0.5}$ scores are listed in Table 2. Using 'clean' sentences in addition to our original ESL data for training always positively affects performance. In this regard, as experimented in (Rei and Yannakoudakis, 2016), training on more data in the same domain is a valid solution for improving the performance of LSTM models. However, when also adding artificially generated data to the training set, we reach higher scores only on 2 out of the 4 benchmarks. It greatly improves the average recall (+11.03), without hurting the precision on FCE and CoNLL14 but affects negatively the precision on AESW and JFLEG.

	FCE	**AESW**	**CoNLL14**	**JFLEG**
System	$F_{0.5}$	$F_{0.5}$	$F_{0.5}$	$F_{0.5}$
LSTM$_{ESL}$	53.49	20.75	38.46	57.82
LSTM$_{ESL+art}$	66.04	21.31	46.27	57.29
LSTM$_{ESL+cor}$	65.08	**27.16**	46.26	**59.52**
LSTM$_{ESL+art+cor}$	**67.16**	21.12	**52.28**	54.64

Table 2: Performance ($F_{0.5}$ scores) of the LSTM models when trained using an additional set of 'clean' sentences (*cor*) where non-SVA errors have been corrected.

8.2 Sensitivity to long-distance dependencies

Next, we want to study how well our models perform when the subjects and verbs are far apart, i.e., when the agreement relation is defined over a long-distance dependency. In order to see how our systems are affected by the distance between the subject and verb, we split the test sets based on different subject–verb distances.

Note, however, that our benchmarks are not annotated with PoS tags and dependency relations. If we binned our test data based on predicted dependencies, the inductive bias of our syntactic parser and the errors it made would bias our evaluation. Instead, we perform our analyses on section 22 and 23 of the Penn Treebank (PTB) dataset (Marcus et al., 1993). The PTB however is not annotated with grammatical errors. We therefore corrupt the sentences by injecting SVA errors, in the same

way we corrupted the BNC (§5.3) to create additional training data.

For each sentence in the PTB, we identify a subject–verb pair, and group the sentences by the subject–verb distance. We then run our models on two versions of each sentence: an unaltered version and a corrupted one, where we have generated an SVA error by corrupting the verb, using the method described earlier (§5.3). This way we can compute the performance of our models as $F_{0.5}$ scores over this dataset. The results are displayed in Figure 2. We can see that the LSTM trained with artificial data performs significantly better on long-distance subject–verb pairs than the LSTM trained only on ESL data. This suggests that training on artificially generated errors also makes our models more robust to this potential source of error.

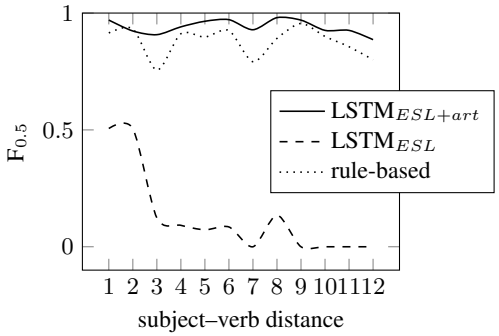

Figure 2: $F_{0.5}$ scores of the systems on the PTB as a function of subject–verb distance.

Note that, in general, there is a substantial gap between the performance of the two LSTM models. This is because one is trained on artificial data – similar to the data we use in our analysis. However, the conclusions are based on the relative differences in performance over long-distance dependencies, and these differences should still be comparable across the two models.

8.3 Sources of error for our rule-based baseline

There are two obvious potential sources of error for our rule-based baseline: sensitivity to errors in the underlying syntactic parsers, and sensitivity to the irregularities of language, e.g., when collective nouns or named entities are subjects, subject–verb agreement cannot always be determined by the PoS tags. We show that the main source of error seems to be irregularities by showing that the underlying syntactic parsers perform relatively well, even in the ESL domain.

Table 3 lists the parsing and tagging performance of our underlying syntactic parsers across three domains: learner data (ESL) and web data (EWT) from the Universal Dependencies (UD) project (Nivre et al., 2017), as well as the newswire data it was trained on (PTB). We only evaluate subject–verb relations, since these are the only ones of interest in this paper. We see that while there is a noticeable out-of-domain drop going from newswire to learner language or web data, the parser is still able to detect subject–verb relations with high precision and recall. This suggests that the vulnerability of our rule-based baseline is primarily a result of linguistic irregularities and exceptions to the implemented rules.

	UD-ESL	UD-EWT	PTB 23
Subject–verb precision	88.47	88.86	91.31
Subject–verb recall	89.37	85.11	89.84
PoS tags accuracy	96.36	93.20	97.79

Table 3: The Stanford PoS Tagger and Dependency Parser's performance on different treebanks. Subject–verb precision/recall relates to subject–verb relations. PoS tag accuracy is only for PoS tags of the subjects and verbs.

8.4 Sensitivity to other linguistic phenomena

Finally, manually reviewing the errors made by the rule-based system, we identified frequent linguistic sources of errors, including relative clauses, conjunctions, ambiguous PoS tags, and collective nouns. We therefore analyze how the LSTMs and the rule-based system are globally sensitive to these potential sources of error. Since our benchmarks are not annotated with PoS and dependency relations, we again use the corrupted PTB sentences (see §8.2).

Many of the examples in which our rule-based baseline fails include *relative clauses* (when the verb is the root of a relative clause) and *conjunctions* (when the subject is a conjunction). A second major cause of failure is ambiguous verbs, i.e., verb forms that can also be nouns (*ambiguous PoS*, e.g., "need", "stop", "point", etc.), and subjects which are singular nouns describing groups of people or things (*collective nouns*, e.g., "team", "family", "staff", etc.). The following examples illustrate these cases (underlined):

(3) a. The church and the cathedral **are** very interesting [...] (*conjunction*)

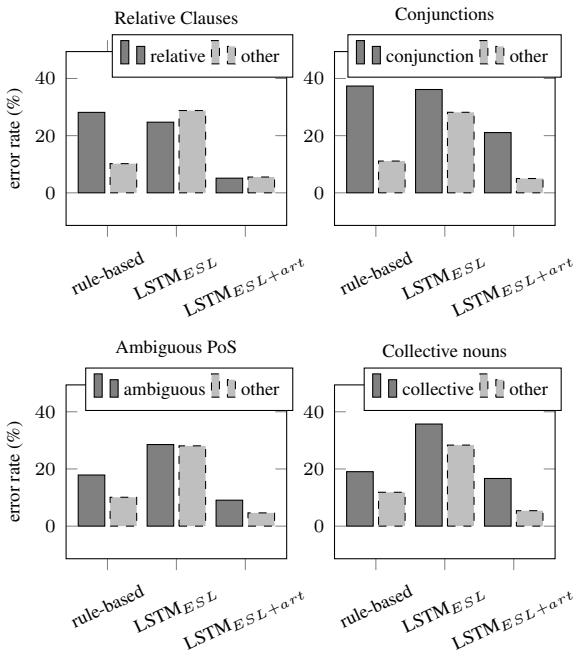

Figure 3: SVA error rates on the PTB data for complex syntactic structures and ambiguous cases.

b. If there is someone who **doesn't** agree with me, he or she [...] (*relative clause*)
c. It is said that the majority of the citizens **has** got a car [...] (*collective noun*)
d. [...] and police officer **walk** around the building as well. (*ambiguous PoS*)

We evaluate our models on the PTB data and report the error rate (the lower the better) on present tense verbs (Figure 3). Overall, results show that all models are negatively affected when they encounter complex syntactic structures and ambiguous cases. Figure 3 also confirms that the rule-based baseline is the most sensitive one to complex structures. Especially in comparison with the LSTM$_{ESL+art}$ model, the rule-based system achieves good scores on verbs which are not part of complex structures, but performs significantly worse on difficult cases. The LSTM$_{ESL}$ model is the worst across almost all cases, while the LSTM$_{ESL+art}$ shows significant improvements over the baselines, in particular for the difficult cases.

9 Conclusion

In this paper, we argue for artificial error generation as an effective approach to learning more robust neural models for subject–verb agreement detection. We demonstrate that error generation is much less sensitive to parsing errors and irregularities than rule-based systems for detecting subject–verb agreement. On the other hand, artificial error generation enables us to utilise much more training data, and therefore can develop more robust neural models for SVA error detection that do not overfit the available, manually annotated training data. Our simple approach to detecting subject–verb agreements achieves a new state of the art on three out of four available benchmarks, and, on average, is better than previous approaches on the task. We show that, in particular, models trained on large volumes of artificially generated errors become more robust to other errors in the surrounding context of SVA, long-distance dependencies, and other challenging linguistic phenomena.

Acknowledgements

This project was supported by Siteimprove and the Innovation Fund of Denmark through an industrial PhD grant. Marek Rei and Helen Yannakoudakis were supported by Cambridge Assessment, University of Cambridge.

References

Øistein E. Andersen, Helen Yannakoudakis, Fiona Barker, and Tim Parish. 2013. Developing and testing a self-assessment and tutoring system. In *Proceedings of BEA 2013*.

BNC-Consortium et al. 2007. The British National Corpus, Version 3. *Distributed by Bodleian Libraries, University of Oxford*.

Chris Brockett, William B Dolan, and Michael Gamon. 2006. Correcting ESL Errors Using Phrasal SMT Techniques. In *Proceedings of COLING-ACL 2006*.

Christopher Bryant and Ted Briscoe. 2018. Language Model Based Grammatical Error Correction without Annotated Training Data. In *Proceedings of BEA 2018*.

Christopher Bryant, Mariano Felice, and Ted Briscoe. 2017. Automatic Annotation and Evaluation of Error Types for Grammatical Error Correction. In *Proceedings of ACL 2017*.

Dongfeng Cai, Yonghua Hu, Xuelei Miao, and Yan Song. 2009. Dependency Grammar Based English Subject-Verb Agreement Evaluation. In *Proceedings of PACLIC 2009*.

Danqi Chen and Christopher Manning. 2014. A Fast and Accurate Dependency Parser using Neural Networks. In *Proceedings of EMNLP 2014*.

Shamil Chollampatt and Hwee Tou Ng. 2018. A Multilayer Convolutional Encoder-Decoder Neural Network for Grammatical Error Correction. In *Proceedings of AAAI 2018*.

Daniel Dahlmeier, Hwee Tou Ng, and Siew Mei Wu. 2013. Building a Large Annotated Corpus of Learner English: The NUS Corpus of Learner English. In *Proceedings of BEA 2013*.

Vidas Daudaravicius, Rafael E. Banchs, Elena Volodina, and Courtney Napoles. 2016. A Report on the Automatic Evaluation of Scientific Writing Shared Task. In *Proceedings of BEA 2016*.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. *arXiv preprint arXiv:1810.04805*.

Mariano Felice and Zheng Yuan. 2014. Generating artificial errors for grammatical error correction. pages 116–126.

Kristina Gulordava, Piotr Bojanowski, Edouard Grave, Tal Linzen, and Marco Baroni. 2018. Colorless Green Recurrent Networks Dream Hierarchically. In *Proceedings of NAACL 2018*.

Sepp Hochreiter and Jürgen Schmidhuber. 1997. Long Short-term Memory. *Neural Computation*, 9.

Sudhanshu Kasewa, Pontus Stenetorp, and Sebastian Riedel. 2018. Wronging a right: Generating better errors to improve grammatical error detection. In *Proceedings of EMNLP 2018*.

Christo Kirov, John Sylak-Glassman, Roger Que, and David Yarowsky. 2016. Very-large Scale Parsing and Normalization of Wiktionary Morphological Paradigms. In *Proceedings of LREC 2016*.

Adhiguna Kuncoro, Chris Dyer, John Hale, Dani Yogatama, Stephen Clark, and Phil Blunsom. 2018. LSTMs Can Learn Syntax-Sensitive Dependencies Well, But Modeling Structure Makes Them Better. In *Proceedings of ACL 2018*.

Guillaume Lample, Miguel Ballesteros, Sandeep Subramanian, Kazuya Kawakami, and Chris Dyer. 2016. Neural architectures for named entity recognition. In *Proceedings of NAACL 2016*.

Claudia Leacock, Martin Chodorow, Michael Gamon, and Joel Tetreault. 2010. Automated Grammatical Error Correction for Language Learners. *Synthesis lectures on human language technologies*, 3(1):1–134.

Tal Linzen, Emmanuel Dupoux, and Yoav Goldberg. 2016. Assessing the Ability of LSTMs to Learn Syntax-Sensitive Dependencies. *Transactions of the Association for Computational Linguistics (TACL)*, 4.

Alexis Manaster-Ramer. 1987. Subject-verb Agreement in Respective Coordinations and Context-freeness. *Computational Linguistics*, 13:64–65.

Mitchell P. Marcus, Mary Ann Marcinkiewicz, and Beatrice Santorini. 1993. Building a Large Annotated Corpus of English: The Penn Treebank. *Computational linguistics*, 19(2):313–330.

Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg S. Corrado, and Jeff Dean. 2013. Distributed Representations of Words and Phrases and their Compositionality. In *Proceedings of NIPS 2013*.

Tomoya Mizumoto, Mamoru Komachi, Masaaki Nagata, and Yuji Matsumoto. 2011. Mining Revision Log of Language Learning SNS for Automated Japanese Error Correction of Second Language Learners. In *Proceedings of IJCNLP 2011*.

Ryo Nagata and Kazuhide Nakatani. 2010. Evaluating performance of grammatical error detection to maximize learning effect. In *Proceedings of COLING 2010*.

Courtney Napoles, Keisuke Sakaguchi, and Joel Tetreault. 2017. JFLEG: A Fluency Corpus and Benchmark for Grammatical Error Correction. In *Proceedings of EACL 2017*.

Hwee Tou Ng, Siew Mei Wu, Ted Briscoe, Christian Hadiwinoto, Raymond Hendy Susanto, and Christopher Bryant. 2014. The CoNLL-2014 Shared Task on Grammatical Error Correction. In *Proceedings of CoNLL 2014: Shared Task*.

Joakim Nivre, Željko Agić, Lars Ahrenberg, Lene Antonsen, Maria Jesus Aranzabe, Masayuki Asahara, Luma Ateyah, Mohammed Attia, Aitziber Atutxa, Liesbeth Augustinus, Elena Badmaeva, Miguel Ballesteros, et al. 2017. Universal dependencies 2.1. LINDAT/CLARIN digital library at the Institute of Formal and Applied Linguistics (ÚFAL), Faculty of Mathematics and Physics, Charles University.

Marek Rei. 2017. Semi-supervised Multitask Learning for Sequence Labeling. *Proceedings of ACL 2017*.

Marek Rei, Mariano Felice, Zheng Yuan, and Ted Briscoe. 2017. Artificial Error Generation with Machine Translation and Syntactic Patterns. In *Proceedings of BEA 2017*.

Marek Rei and Helen Yannakoudakis. 2016. Compositional Sequence Labeling Models for Error Detection in Learner Writing. In *Proceedings of ACL 2016*.

Marek Rei and Helen Yannakoudakis. 2017. Auxiliary Objectives for Neural Error Detection Models. In *Proceedings of BEA 2017*.

Alla Rozovskaya and Dan Roth. 2011. Algorithm Selection and Model Adaptation for ESL Correction Tasks. In *Proceedings of ACL 2011*.

Nitish Srivastava, Geoffrey Hinton, Alex Krizhevsky, Ilya Sutskever, and Ruslan Salakhutdinov. 2014. Dropout: A Simple Way to Prevent Neural Networks from Overfitting. *The Journal of Machine Learning Research*, 15(1):1929–1958.

Guihua Sun, Gao Cong, Xiaohua Liu, Chin-Yew Lin, and Ming Zhou. 2007. Mining Sequential Patterns and Tree Patterns to Detect Erroneous Sentences. In *Proceedings of AAAI 2007*.

Kristina Toutanova, Dan Klein, Christopher D. Manning, and Yoram Singer. 2003. Feature-Rich Part-of-Speech Tagging with a Cyclic Dependency Network. In *Proceedings of NAACL 2003*.

Yuzhu Wang, Hai Zhao, and Dan Shi. 2015. A Light Rule-based Approach to English Subject-Verb Agreement Errors on the Third Person Singular Forms. In *Proceedings of PACLIC 2015*.

Ziang Xie, Guillaume Genthial, Stanley Xie, Andrew Ng, and Dan Jurafsky. 2018. Noising and Denoising Natural Language: Diverse Backtranslation for Grammar Correction. In *Proceedings of NAACL 2018*.

Helen Yannakoudakis, Ted Briscoe, and Ben Medlock. 2011. A New Dataset and Method for Automatically Grading ESOL Texts. In *Proceedings of ACL 2011*.

Helen Yannakoudakis, Marek Rei, Øistein E Andersen, and Zheng Yuan. 2017. Neural Sequence-Labelling Models for Grammatical Error Correction. In *Proceedings of EMNLP 2017*.

Matthew D. Zeiler. 2012. ADADELTA: An Adaptive Learning Rate Method. *arXiv preprint arXiv:1212.5701*.